**GROCERY RETAIL FORECASTING USING TIME SERIES FORECASTING METHODS**

Sushil Ram Achamwad, Udaiveer singh Chauhan, Melvyn Thomas Daniel Sagayaraj,

Sagar Pradhan, Matthew A. Lanham

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

sa[chamwa@purdue.edu](mailto:chamwa@purdue.edu); [uchauhan@purdue.edu](mailto:uchauhan@purdue.edu);

[danie149@purdue.edu](mailto:danie149@purdue.edu); [pradhan4@purdue.edu](mailto:pradhan4@purdue.edu); [lanhamm@purdue.edu](mailto:lanhamm@purdue.edu)

## **ABSTRACT**

We built various demand forecasting models to predict product demand for grocery items using R Packages. It is a timeseries analysis problem and there are various methodologies to solve these kinds of problems. Time series forecasting is a technique for the prediction of events through a sequence of time and it is technique to predict future events by [analyzing](https://searchbusinessanalytics.techtarget.com/definition/predictive-analytics) the trends of the past, on the assumption that future trends will hold similar to historical trends. The methods we’ve employed our model are ARIMA, XGBoost , ETS (Exponential Smoothing State Space Model and Prophet . comparisons along the following dimensions: 1) predictive performance, 2) scalability 3 ) ease of use scalability .Tool wise we have used R Studio to code the model. On the basing weightage on the above mentioned comparisons we chose the best time series modelling techniques and predicted the demand values for the next two weeks.

Keywords: ARIMA, XGBoost , ETS (Exponential Smoothing State Space Model ,Prophet , Timeseries Analysis

**INTRODUCTION**

There are lot of benefits for effectively predicting sales forecasts.(1) Efficient supply chain scheduling where Forecasting the amount of sales and when they are likely to occur, would help better scheduling production, warehousing and shipping. (2)Better Labor Management where Anticipating demand means knowing when to increase staff and other resources to keep operations running smoothly during peak periods. (3) Adequate Cash Flow: Knowing the peaks and valleys of demand would help in better management of cash flow, ensuring there is enough money and bills are paid on time. (4)More Accurate Budgeting: The more accurately you can forecast demand, including the timing of your sales, the more accurate you can be with budgeting. Overstocking can be prevented at the Inventory SKU’s thereby prevention of wastage loss.

Here we are Predicting future demand for different items in a grocery store. [Corporación Favorita](http://www.corporacionfavorita.com/) is a large Ecuadorian-based grocery retailer which operates hundreds of supermarkets, with over 200,000 different products on their shelves. They want to know how much they should stock up in order to prevent overstocking . They want to solve this using an analytical approach. It is a timeseries analytics problem and there are various methodologies to solve these kinds of problems. The client, [Corporación Favorita](http://www.corporacionfavorita.com/), has come to an agreement for allowing the usage of their data for analyzing and coming up with a solution.

Gartner in it’s publication “Demand Forecasting Leads the List of Challenges Impacting Customer Service Across Industries”(2016). They state that demand forecast across all customer facing industries are important for the business.

# Forbes in it’s publication “Forecasting And Budgeting Can Improve Your Company's Fiscal Performance

”(2017). They state that about 70 % percentage the top successful companies have invested in business forecasting part.

A Wall Street Journal article, “Retailers Rethink Inventory Strategies (Ziobro, 2016)” mentions how Home Depot is trying to minimize its inventory at stores. We see that there is an increasing need for demand forecasting techniques that can accurately predict the demand for each item for each day for every store. This need is being fulfilled in some companies by using open source data science tools whereas few other firms use in-house commercial platforms.

In this paper, we try to evaluate the predictive model performance of models using open-source data science tools like R and Business Intelligence Tool tableau to predict demand for thousands of products on a store level for a Kaggle competition dataset. The performance metric which we have use does not stop with model accuracy. We posit that the real value of a model to a business is a composite of (1) predictive model accuracy, (2) scalability and (3) ease of use.

Our paper structure is as follows . Literature review is done to see which forecast method is the successful one to proceed with . We will be discussing the data set used in our study. Next, we discuss the methodology/design we implemented and discuss the models we investigated. Lastly, we present our results, discuss our conclusions, and how we plan to extend this research.

## **LITERATURE REVIEW**

A strategic goal for any retailer is to maintain the stocks of all products so as not to run out of stock when the customers need it. Identifying the variables that affect the sales of a products and ensuring the stocks are maintained to serve the best potential-revenue generating accounts is critical to a company’s revenue growth. An analytical challenge is to predict the potential sales of an item given the trend of sales of that product over the past few years in different conditions.

We started our search for the optimal prediction model for forecasting by looking at past research done in demand forecasting using different machine learning algorithms. This exercise gave us an understanding of different machine learning models that could be used for forecasting and their respective strengths and weaknesses. We also looked at the various measures frequently employed to compare their performances.

Sales forecasting is a complex process for several reasons. There are multiple stages involved, each stage has several participants (from the buyer and seller’s side) who may not necessarily have the same objectives and interests. Sales forecasts are a critical cog in making managerial decisions and incorrect forecasting can lead to wasting of resources (Bohanec, Robnik-Šikonja, & Borštnar, 2017).

Before applying a prediction model to any analytical problem, the user must trust that the model will fulfill its purpose, this trust can be built based on the performance as well as the interpretability of the model. Hence, while more sophisticated models like neural networks and support vector machines (SVMs) might generate a better overall predictive model, they simply lack the interpretability of simpler models like linear regression and decision trees (Caruana & Niculescu-Mizil, 2006).

Previous research on demand forecasting has traditionally used a methodology called Autoregressive Integrated Moving Average (ARIMA). This methodology has been applied to studies of traffic flow (Williams, 2003) and international travel demand (Lim, 2002). Lim’s paper analyzed stationary and non-stationary international tourism time-series data by formally testing for the presence of unit roots and seasonal unit roots prior to estimation, model selection, and forecasting. They used mean absolute percentage error (MAPE) and root mean squared error (RMSE) as measures of forecast accuracy. This paper showed that by comparing the RMSE’s, lower post-sample forecast errors were obtained when time-series methods such as the Box–Jenkins ARIMA and seasonal ARIMA models were used.

An accurate sales forecasting system is an efficient way to achieve increased profits and reduced costs. Poor forecasting may result in redundant or insufficient stock, which will directly affect the revenue and competitive position. Some of the international studies on retail sales forecasting attempt to select the optimal forecasting model by comparing forecasts from single artificial neural network (ANN) models with one or two traditional methods such as the exponential smoothing, moving average (MA), autoregressive and integrated moving average (ARIMA), seasonal ARIMA (SARIMA) and generalized autoregressive conditional heteroscedastic (GARCH) models (Aye & Balcilar, 2014).

The choice of appropriate forecasting methods and sources of information play an important role when it comes to forecasting demand in the retail food stores, and so does the identification of all the factors that could significantly affect future demand of customers. Due to the nature of demand on consumer markets, as well as the availability of accurate data on customers’ past sales (POS data), it is recommended to use mainly quantitative methods based on the time series analysis (Patak, 2015).

To generate a better rounded prediction model, (Qu and Zhang, 2017) included several external factors in addition to the price and sales of an item. Factors like the temperature, fuel price, unemployment rate and the relative price of an item were considered for model generation. Sales data for 2.5 years was used for the study. The study deployed a random forest with regression trees method for the analysis by employing the rpart() method to understand variable importance. They were able to conclude that their adopted forecasting method performed better as compared to other traditional forecasting techniques such as Least Square regression and Principle components regression by taking the Mean Absolute Percentage Error (MAPE) as the standard for measurement.

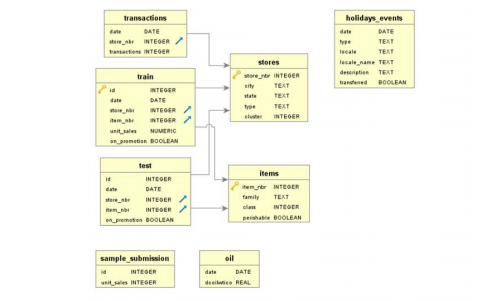
Another factor that plays an important role in the sales of a product is whether the item is on some kind of promotion or available for a discounted price. (van Donselaar & Peter, 2016) developed a forecasting model for the demand of perishable products in retail stores to better understand the impact of price discounts. The success of the promotions is determined by a Lift factor (sales after promotions as compared to baseline sales) which is also the dependent variable in forecasting models further developed in the study. It was concluded that Large part of the variation in the effect of promotions was explained by the selected independent variables. To understand multicollinearity VIF statistic was used and results concluded that multicollinearity didn’t pose any concern for these models.

Usually in case one wants to forecast the time series data in R, the ‘forecast’ package is typically employed with models like ARIMA. In 2017, a core data science team at Facebook released ‘Prophet’, an R package which utilizes a Bayesian based curve fitting method to forecast the time series data. The major advantage of Prophet is that it doesn’t require much prior knowledge or experience of forecasting time series data as it automatically finds seasonal trends beneath the data, offering a set of easy to understand parameters. Using the Prophet package is relatively easy and allows users to get reasonably good results on par with those produced by experts (Hayashi, 2017).

|  |  |  |
| --- | --- | --- |
| **Studies** | **Motivation for the Research** | **Result of the Research** |
| Caruana & Niculescu-Mizil (2006) | To present a large-scale empirical comparison between various supervised learning methods | * Learning methods like boosting, random forest and SVMs achieve excellent performance * Calibrated boosted trees are the best overall learning algorithm |
| Williams, B. M., & Hoel, L. A. (2003)  Lim, C., & McAleer, M. (2002) | To study the traditional methods of demand forecasting | * ARIMA models can give good accuracy * May cause problems in the initial model selection as they are based on heuristic selection of parameters * Can be time-consuming if many time series observations are to be analyzed * RMSE and MAPE are widely used to measure forecast performance |
| Aye & Balcilar (2014) | To forecast aggregate retail sales in the nation of South Africa | * Models with seasonal dummy variables produce better forecasts than the full seasonal models * Difficult to identify a single model that outperforms all others. |
| Ramos & Santos (2014) | To evaluate the performance of state space and ARIMA models for consumer retail forecasting | * Performance of both ARIMA and State Space better when doing multi-step forecasts than single-step |
| Qu & Zhang (2017) | To predict the demand and optimize the prices for semi-luxury supermarket segment | * Random forest with regression trees method performs the best in this scenario |
| van Donselaar & Peter (2016) | To analyze the effects of promotions on the demand of perishable items and predict the demand | * Large part of the variation in sales can be explained by the selected independent variables * Multicollinearity didn’t pose any concern for these models |
| Guolin Ke, Qi Meng (2017) | To explore high performance GBM methods for data forecasting | * LGBM models can significantly outperform XGBoost and SCG in terms of computational speed and memory consumption |
| Gur Ali & Sayn (2009) | To forecast the demand in an SKU-extensive store in the presence of promotions | * Simple time-series techniques perform very well in absence of promotions * For periods with promotions, regression trees with explicit features improve accuracy substantially. |
| Hayashi (2017) | To compare the performance of ARIMA and Prophet for forecasting time series data | * Prophet performs better than conventional ARIMA * ARIMA model with fine-tuned parameters performed better than Prophet |

## **DATA**

The data used in this research is from the Kaggle competition which aims to forecast demand for millions of items at a store and day level for a South American grocery chain. (https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data). The data is provided in different tables named train, test, stores (store related data), items (merchandise data), transaction, oil (oil prices can be a good predictor of sales as Ecuador is an oil-dependent economy), and holidays events (holiday & major event related data). Table 2 provides a summary of all the important columns is given below along with the relations between each data table provided.



## Figure 1: Dataset Information

## **METHODOLOGY**

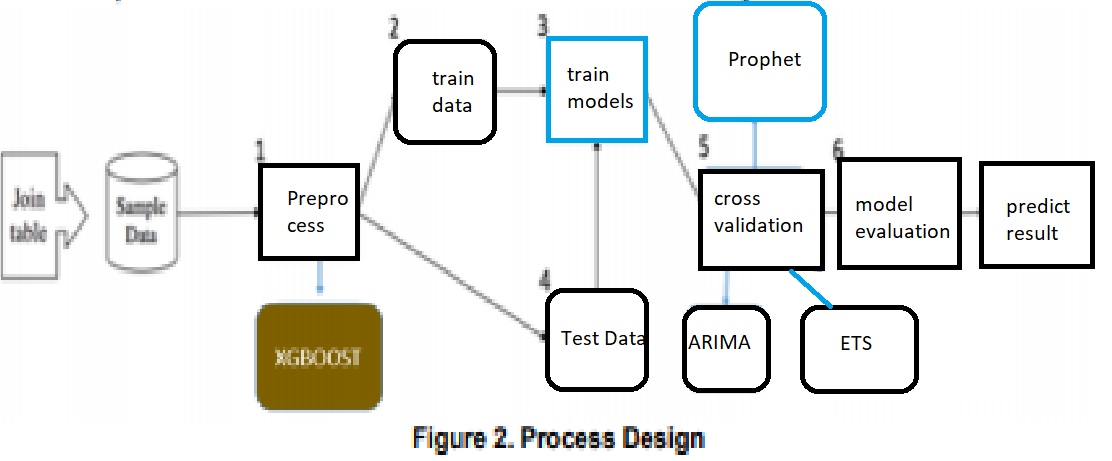
**Business Understanding:**

We started with understanding the business’s objectives of the problem and its application. We understood that in addition to the usual trend in shopping of items, there could be external factors (e.g. Oil Price, Holidays) which could affect the demand. Further, we understood the rationale of high weight given to perishable goods and its impact on the business. As we understood more, it was clear to use that it was a short-term (15th days) demand forecasting required at a granular level. We prepared a preliminary strategy by starting our research more into various modeling technique applicable to time-series and best practices and tooling required in dealing with the BigData (100 million rows of training). Through Literature review, we were mindful of the fact that some model-output (e.g. PCA, Clustering) would serve as input for other models and for this reason, it was included in the strategy to try out various model (with moving average features).

**Data Understanding:** We sourced flat data files and did some preprocessing to do EDA (Exploratory Data Analysis) exercise. Then, with a connection to Tableau, we collected the basic facts about the data and studied the distribution of all the key variables.

**Modelling overview:**

Various models like ARIMA, XGBOOST, ETS and Prophet were applied and the predicted values of test set and trainset were validated against each other. Predictions over a random time slice of 16 day horizon were generated using the below given methods, for a selected store and a selected item, and the RMSE values were calculated for those predicted values.



## **MODEL(s)**

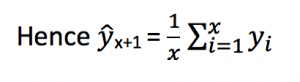
**FORECASTING MODELS**

**ARIMA**

This is one of the oldest and most widely used methods of demand forecasting. In this method, the average sales of the previous days are used as the predictor for the sales of the next day. It is simple and gives good accuracy when done on a short-term horizon. However, it is not likely to predict well for a longer-duration span as it is not generalizing the trend mere following the past behavior with auto-regressive components.

Here, we used arima.fit which takes care of AIC BIC parameters,  AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values are estimators to compare models. The lower these values, the better is the model.

Here the values are predicted based on the average of the previous observations



**XGBOOST**

Extreme gradient Boosting Model(LGBM) is a fast variant method in the class of tree-based boosting algorithm. XGBoost is also known as ‘**regularized boosting**‘ technique. GBoost is also known as ‘**regularized boosting**‘ technique. We chose XGBoost for its wide range of tuning parameters that can be implied to optimize, and it comes with a built-in cross validation model.

Cons: XGBoost because it’s computationally expensive, takes a while to come up with an optimal solution.

obj(θ)=L(θ)+Ω(θ)

L(θ)=∑i(yi−y^i)2

L(θ)=∑i[yiln(1+e−y^i)+(1−yi)ln(1+ey^i)]

**Tuning parameters**

These are the tuning parameters that were used.

xgbGrid <- expand.grid(nrounds = 500 max\_depth = 4, eta = .05, gamma = 0, colsample\_bytree = .5, min\_child\_weight = 1, subsample = 1), nrounds = 500.

maxdepth = The maximum depth of a tree

eta = analogous to learning rate, Makes the model more robust by shrinking the weights on each step

Gamma: A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split. Makes the algorithm conservative. The values can vary depending on the loss function and should be tuned.

colsample\_bytree: Similar to max\_features in GBM. Denotes the fraction of columns to be randomly samples for each tree.

subsample [default=1]: Same as the subsample of GBM. Denotes the fraction of observations to be randomly samples for each tree. Lower values make the algorithm more conservative and prevents overfitting but too small values might lead to under-fitting.

min\_child\_weight [default=1]: Defines the minimum sum of weights of all observations required in a child. This is similar to min\_child\_leaf in GBM but not exactly. This refers to min “sum of weights” of observations while GBM has min “number of observations”. Used to control over-fitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample selected for a tree. Too high values can lead to under-fitting hence, it should be tuned using CV.

**ETS (Exponential Smoothing)**

This method gives more significance to recent observations, and it is used to predict forecast for a season ahead. It neglects the ups and downs associated with random variation which on a graph shows you a smoother line or curve. This it does not take into account the intricate changes caused by different factors within a cycle, gives an overview of the trends happening in the season.

Forecast equation yt+1|t=ℓt

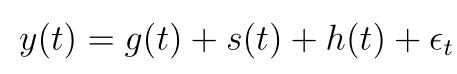
Smoothing equation ℓt=αyt+(1−α)ℓt−1,

Error term: et=yt−ℓt−1=yt−^yt|t−1et=yt−ℓt−1=yt−y^t|t−1 is the residual at time t.

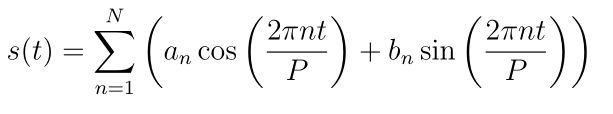
**Prophet**

Prophet was developed by the team at facebook, and it is one of the best methods applied in demand forecasting when factors like seasonal changes, holidays are taken into consideration as it has a specific functional part assigned to each of these changes in its algorithm.

We use a decomposable time series model with three main model components: trend, seasonality, and holidays. They are combined in the following equation:



Here, the seasonality is presented by the equation



* **g(t)**: piecewise linear or logistic growth curve for modelling non-periodic changes in time series
* **s(t)**: periodic changes (e.g. weekly/yearly seasonality)
* **h(t)**: effects of holidays (user provided) with irregular schedules
* **εt**: error term accounts for any unusual changes not accommodated by the model

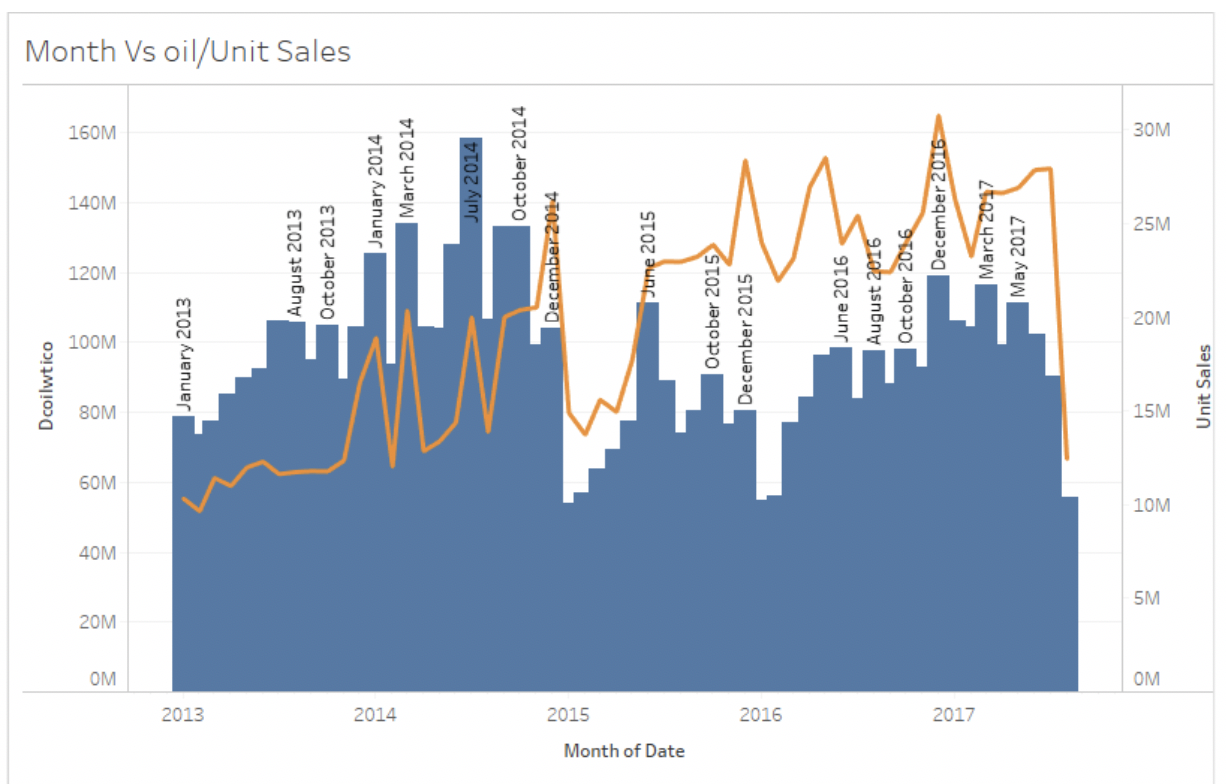
Using time as a regressor, Prophet is trying to fit several linear and non linear functions of time as components. Modeling seasonality as an additive component is the same approach taken by exponential smoothing in [Holt-Winters technique](https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/) . We are, in effect, framing the forecasting problem as a curve-fitting exercise rather than looking explicitly at the time based dependence of each observation within a time series.

Here, as our data problem perfectly aligns with the techniques applied in prophet, like the, holidays, seasonal changes and linear changes, are all taken into account in this method.

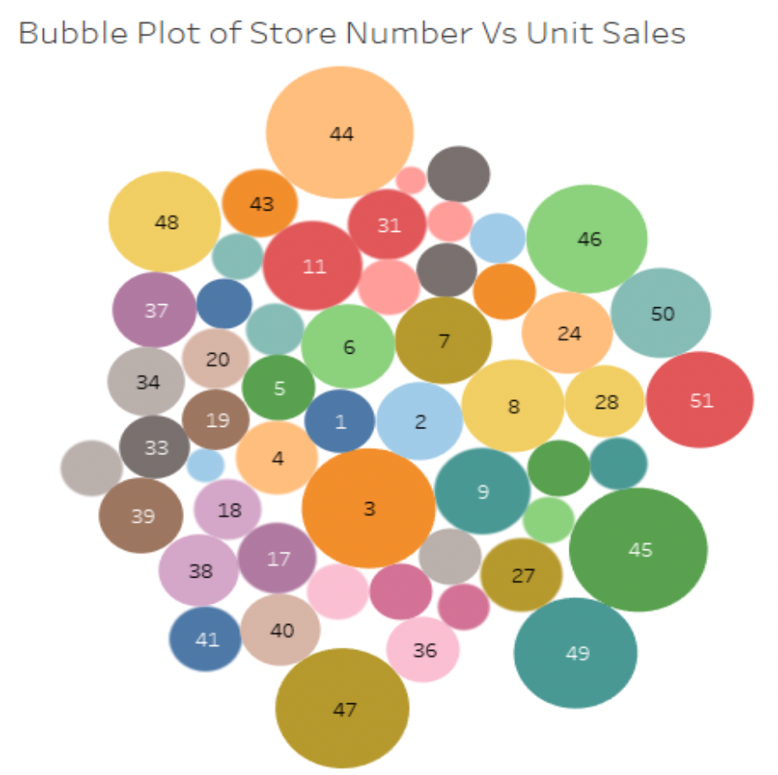
## **RESULTS**

**Descriptive Analytics**

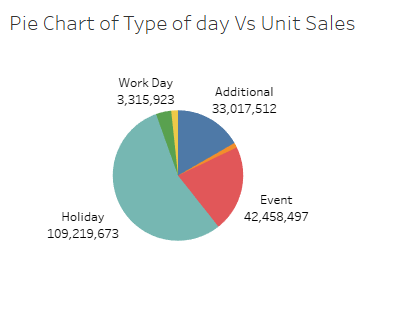
We first perform descriptive analytics on the sample of data we have taken to discern any interesting trends. We plot the total unit sales across all stores with a variety of variables. The plots we got using Tableau are as follows:



Plot 1. Total oil price vs Total units sales through the years



Plot 2: Bubble plot showing store number on the basis of unit sales



Plot 3: Pie chart showing the total unit sales depending in the type of day

We see from the first plot that there seems to be a near inverse relationship between the total oil price and the total unit sales throughout the months. Whenever the oil sales are high, sales tend to be low and vice-versa. The second plot gives us an idea about the stores that tend to sell a greater number of units. Clearly, stores 3, 44, 45 and 47 seem to sell the greatest number of items in the time duration. The third plot gives us an idea about the sum of sales of items on each type of day. Holidays tend to be the most attractive type of day for the grocery stores as most items sell during these days.

However, as we have nearly 221,400 store-item combinations, and the sales for each of these store-items is essentially a unique time series, visually inspecting all the features is not useful in this case.

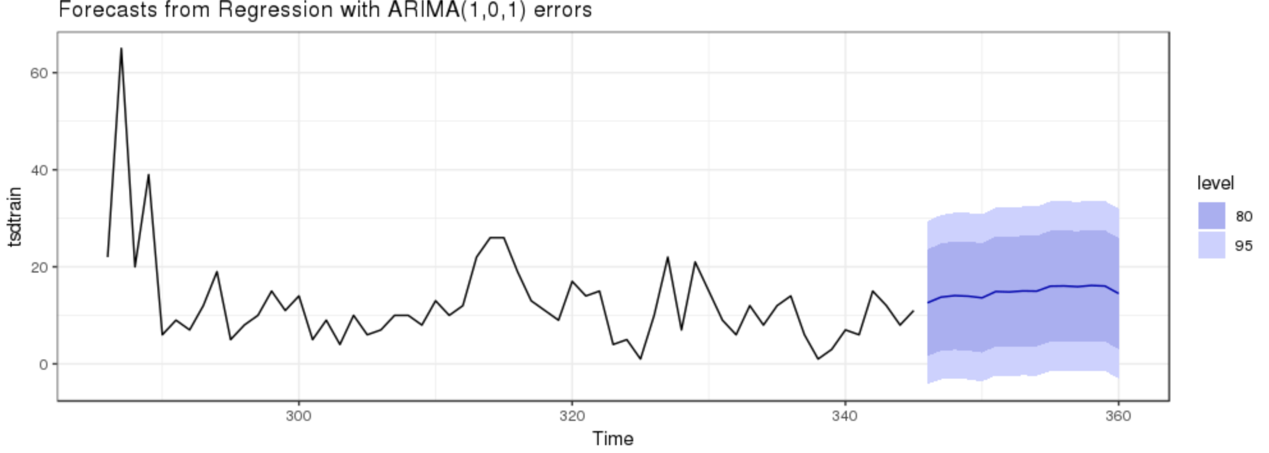
**Forecasting models**

Considering the volume of data provided (125 million observations) and existing computing capacity, only single year of observations (37 Million observations) from 1st Aug 2016 to 31st July 2017 are used for analysis. The objective of this study is to predict no of unit sales for 4100 items sold through 54 stores at various locations from 16th Aug 2017 to 31st Aug 2017.

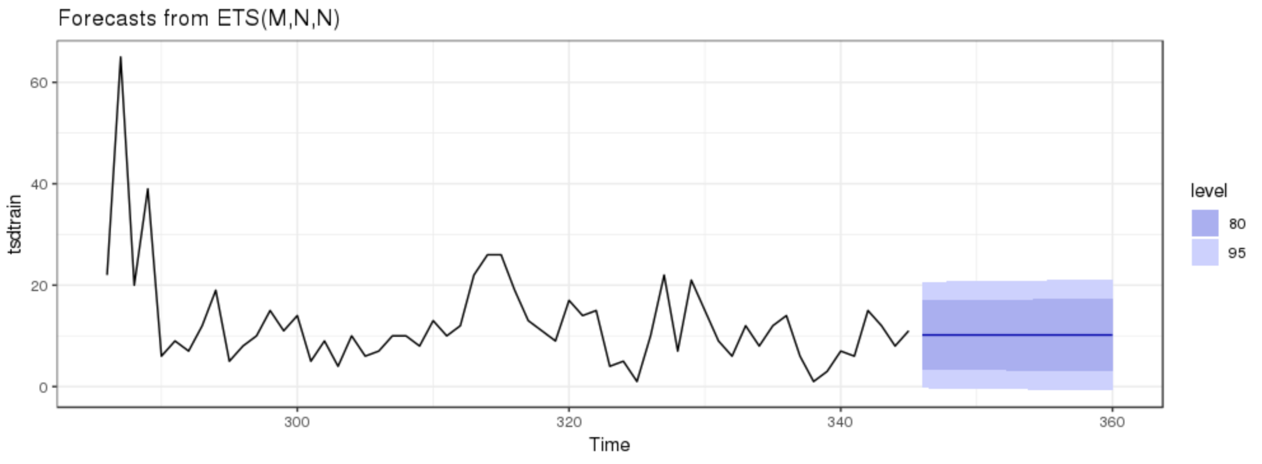
Four different forecasting models were developed for predictions namely, XG Boost, ARIMA, ETS (Exponential Smoothing Test space model) and PROPHET. For assessing the accuracy of each of these models the training data was split into training and testing data set.

Observations from 1st Aug 2016 to 15th July 2017 were used for training and observations from 16th Aug 2017 to 31st Aug 2017 were used for testing. The prediction models were developed at item and store level meaning, for each combination of an item at a store, one separate model was developed to predict its unit sales. For this analysis store number 40 and item number 314393 is selected for this analysis. With this approach, store and item specific information such as location of store, family of items does not affect the model performance as all the training observations for that model would have same values of these variables. 5 fold cross validation was used for assessing the performance of XG Boost models. Multi variate ARIMA model was developed using independent variables such as “Day”, “Weekday”, “Onpromotion”, “Holiday”.

ETS and PROPHET models did not consider any effect of other regressors than time. Following forecasting trends were estimated using selected models:

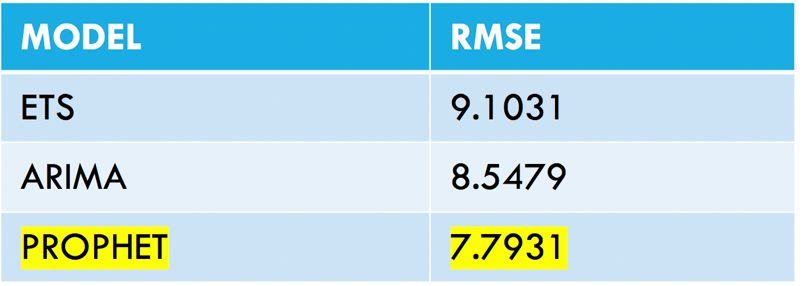


**Unit sales forecasting trend: ARIMA**



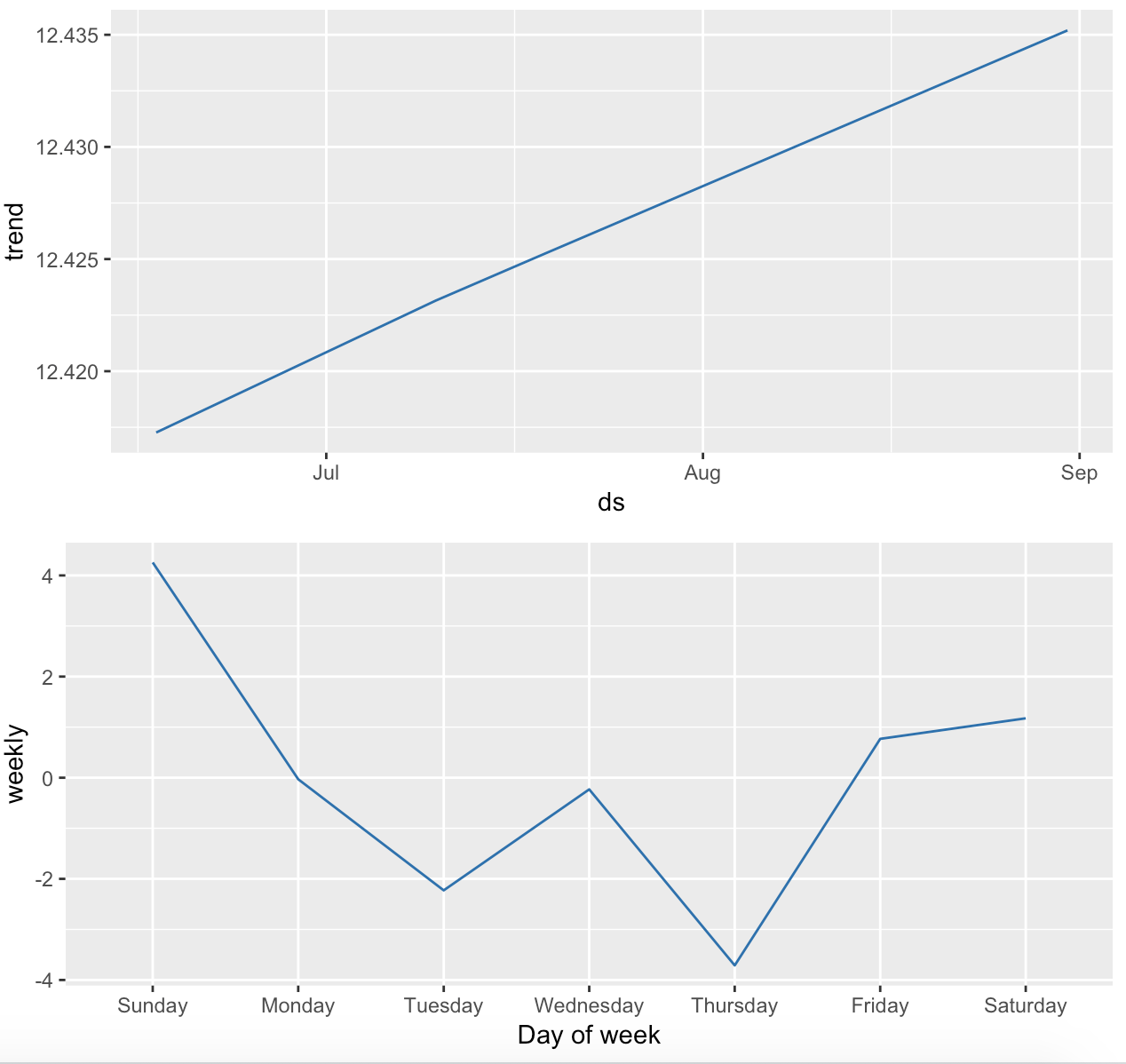
**Average Unit sales forecasting trend: ETS**

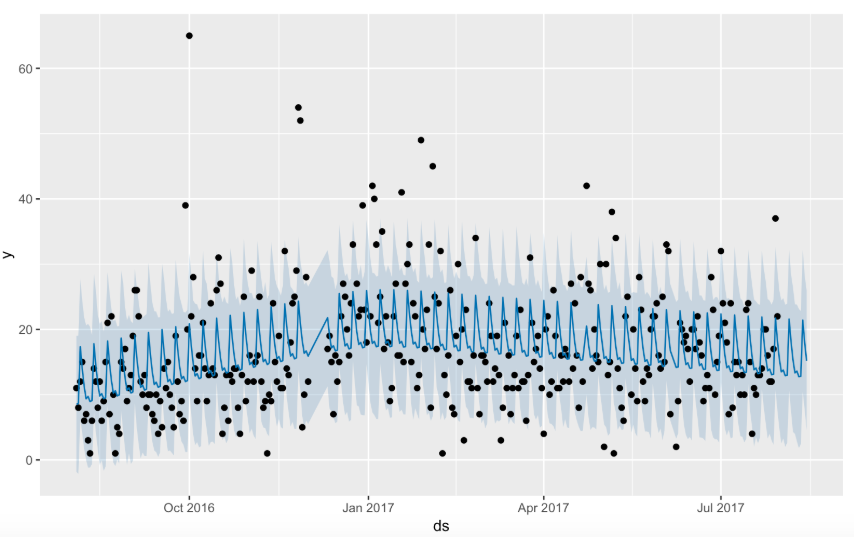
**Best Model Selection:**



RMSE (Root Mean Squared Error) value is used to compare the performance of selected models.

The RMSE value for ARIMA decreased after including other independent regressors from 9.63 to 8.54. But even then, PROPHET performed better than all other models. As shown below, the PROPHET model significantly captures the seasonality in sales trends.





Forecasted Trends – Prophet Model

## **CONCLUSIONS**

A small increase in predictive accuracy can help firms save substantial amount of inventory costs while maintaining acceptable service levels. This was seen with the model Prophet when compared to the other models. The RMSE values were low compared to the other models. As the results of this study demonstrate, machine learning techniques can help in improving the forecast accuracies to a certain extent, in our study, as our research was computationally limited provided the resources, given we had 125 million instances of data. However, with ever increasing technological innovations, future analysts can explore more input features related to intermittent demand prediction and the demand prediction would get extremely accurate over time.

More better model/algorithms can be made, which would take into account the transient changes more readily within a seasonal time forecast which follow a trend, like the one done in prophet where it takes into account the effect of holidays, and other intermittent changes within a trend. More better ways of calculating or predicting these changes would help in better forecasting the timeseries prediction.

## **REFERENCES**

G.C. Aye, M. Balcilar, R. Gupta, A. Majumdar. **Forecasting aggregate retail sales: the case of South Africa**. Int. J. Prod. Econ., 160 (2015), pp. 66-79.

Bohanec, M., Robnik-Šikonja, M., & Borštnar, M. K. (2017). Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting. Organizacija 50(3).

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *23rd International Conference on Machine Learning.* Pittsburg, PA.

Guolin Ke, Q. M. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

Hideaki Hayashi (2017). *Is Prophet Really Better than ARIMA for Forecasting Time Series Data?* Learn Data Science (blog.exploratory.io). https://blog.exploratory.io/is-prophet-better-than-arima-for-forecasting-time-series-fa9ae08a5851

Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics.* John Wiley & Sons, Inc.

Lim, C. &. (2002). Time series forecasts of international travel demand for Australia. *Tourism Management, 23(4)*, 389-396.

Ö.G. Ali, S. Sayn, T. van Woensel, J. Fransoo. ***SKU demand forecasting in the presence of promotions.*** Expert Systems with Applications, 36 (10) (2009), pp. 12340-12348.

P. Ramos, N. Santos, R. Rebelo. ***Performance of state space and ARIMA models for consumer retail sales forecasting***. Robot Comput. Integr. Manuf., 34 (2015), pp. 151-163.

T.Qua, J.H.Zhangb, Felix T.S.Chanc, R.S.Srivastavad, M.K.Tiwarie, Woo-YongPark. *Demand prediction and price optimization for semi-luxury supermarket segment*. Computers & Industrial Engineering Volume 113, November 2017, Pages 91-102.

Van Donselaar, K. H, J. Peters, A. De Jong, R.A.C.M. Broekmeulen. ***Analysis and forecasting of demand during promotions for perishable items*.** Int. J. Prod. Econ., 172 (2016), pp. 65-75.

Vladimira Vlckova, Michal Patak (2010). Role of demand planning in business process management. *6th International Scientific Conference*.

Williams, B. M. (2003). *Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results.* Journal of Transportation Engineering, 129(6), 664-672.